Machine Learning (2019) 108:475–488
https://doi.org/10.1007/s10994-018-5752-y

Majority vote ensembles of conformal predictors

Giovanni Cherubin[1]

Received: 7 February 2018 / Accepted: 1 August 2018 / Published online: 31 August 2018

Abstract
We study majority vote ensembles of ε-valid conformal predictors (CP). We show that the prediction set Γ^{η} produced as the majority vote among the prediction sets Γ_i^{ε} of k independent ε-valid CPs is also valid, for some significance level η; we provide a method to compute ε to achieve a desired η. We further indicate an error upper bound for an ensemble of correlated CPs, and derive a value ε for which such an ensemble guarantees η conservative validity. We evaluate empirically our findings, and compare them with alternative strategies for combining CPs' predictions.

Keywords Conformal prediction · Ensembles · Majority vote · Error bounds

1 Introduction

Conformal predictors (CP) are wrappers around machine learning (ML) classifiers (here called *nonconformity measures*),[1] equipping them with the validity property: for a test object $x \in X$ with label $y \in Y$, and for a chosen significance level $\varepsilon \in [0, 1]$, a CP predicts a set $\Gamma^{\varepsilon} \subseteq Y$ of candidate labels (*prediction set*); such prediction set is ε-valid, in that it guarantees that $Pr\,(y \notin \Gamma^{\varepsilon}) = \varepsilon$. To evaluate the tightness of a CP's predictions, an *efficiency* criterion (e.g., average size of Γ^{ε}) is adopted in applications (Vovk et al. 2005, 2016).

Internally, a CP uses the nonconformity measure to perform a randomness test for a test object x, given a training set of examples $(x_i, y_i)_{i=1}^{n}$, which outputs a p-value for each possible label $\hat{y} \in Y$ for the hypothesis: "$(x, \hat{y})$ belongs to the same distribution as the training examples". The p-value assigned to each candidate label $\hat{y}$ is then thresholded by ε to decide whether to accept the hypothesis, and thus to include $\hat{y}$ in the prediction set Γ^{ε}.

In ML applications, it is generally a good idea to combine the predictions of many classifiers (i.e., to form an ensemble). This tends to lead to improvements when the ensemble's classifiers are either independent or negatively correlated; however, benefits sometimes exist when classifiers are positively correlated (Kuncheva et al. 2003). In terms of CP, an

[1] More precisely, a nonconformity measure is a scoring function (e.g, a classifier with probabilistic output).

Editors: Alexander Gammerman, Vladimir Vovk, Henrik Bostrom, and Lars Carlsson.

✉ Giovanni Cherubin
Giovanni.Cherubin.2013@live.rhul.ac.uk

[1] Royal Holloway University of London, Egham, UK

ensemble would ideally: (i) maintain the validity property, and (ii) improve the predictions' efficiency (Toccaceli and Gammerman 2017).

Previous research proposed CP ensembles that work by combining p-values (e.g., Toccaceli and Gammerman 2017; Linusson et al. 2017). Of them, only the method by Toccaceli and Gammerman (2017) guaranteed validity, in the sense that the combined p-values would produce η-valid prediction sets when thresholded by η; even so, validity of their method only held if assuming independence between the ensemble's CPs.

In this paper, we construct an η-valid CP ensemble by combining the prediction sets of ε-valid CPs via majority vote: we include a label $y \in Y$ in the prediction set Γ^η of the ensemble if y is contained in a majority of the prediction sets Γ_i^ε of the CPs. The significance level η of the ensemble depends on the one of the individual CPs, ε. We determine two strategies for computing the significance level ε required to guarantee an η-valid ensemble: the *I-method*, and the *C-method*. The *I-method* guarantees exact validity, under the assumption that the ensemble's CPs are independent (Sect. 3). The *C-method* guarantees conservative validity (i.e., probability of error is at most η), even when CPs are correlated (Sect. 4).

2 Preliminaries

In an on-line setting, we observe a training sequence of objects and respective labels $(x_1, y_1), \dots, (x_n, y_n)$, and a test object x with label y. We require that examples $(x_1, y_1), \dots, (x_n, y_n), (x, y)$ are exchangeable, sampled from some distribution over a space $X \times Y$, with finite Y.

A CP, $C^{A,\varepsilon} : (X, Y)^* \times X \mapsto \mathcal{P}(Y)$, with nonconformity measure A and significance level $\varepsilon \in [0, 1]$, is an algorithm taking as input a training set of examples $(x_i, y_i)_{i=1}^n$ and a new object $x \in X$, and returning a prediction set $\Gamma^\varepsilon \subseteq Y$ of candidate labels for x (Vovk et al. 2005). We describe the CP algorithm into details in "Appendix". Because the formal part of this paper is agnostic of the nonconformity measure, we will omit A from our notation, and refer to a CP simply with C^ε. In empirical evaluation (Sect. 5), we will specify the nonconformity measure associated with a CP when required.

The following result holds for a CP.

Theorem 1 [CP validity (Vovk et al. 2005)] *Let C^ε be a CP, for some significance level $\varepsilon \in [0, 1]$. Let*

$$\Gamma^\varepsilon := C^\varepsilon(((x_1, y_1), \dots, (x_n, y_n)), x)$$

be the prediction set associated with a test object x with true label y, given $n > 0$ training examples (x_i, y_i). Then C^ε is ε-valid, in that it guarantees:

$$Pr\left(y \notin \Gamma^\varepsilon\right) = \varepsilon \quad .$$

The validity defined in this theorem is an exact validity. There exists an alternative formulation of the CP algorithm, *deterministic* CP, which gives conservative validity, i.e., $Pr\,(y \notin \Gamma^\varepsilon) \leq \varepsilon$. While in this paper we do not treat deterministic[2] CPs, we will later make use of the notion of conservative validity. We remark that results we obtain here for CPs hold similarly for deterministic CPs.

Because the validity of a CP is guaranteed, its performances are measured by its *efficiency*, which indicates the tightness of its predictions. In experiments, we will use the *N criterion*,

[2] CPs considered in this paper are formally known as *smoothed* CPs.

a widely used efficiency criterion that is defined as the average size of the prediction set, $|\Gamma^{\varepsilon}|$ (Vovk et al. 2016).

3 Majority vote CP ensemble

Construct an odd number k of CPs, $C_1^{\varepsilon}, \dots, C_k^{\varepsilon}$, for some significance level ε, whose value we will specify later, and use them to predict a test object x. From their prediction sets $\Gamma_1^{\varepsilon}, \dots, \Gamma_k^{\varepsilon}$, with $\Gamma_i^{\varepsilon} = C_i^{\varepsilon}(((x_1, y_1), \dots, (x_n, y_n)), x)$, we define the majority vote prediction set (i.e., their ensemble prediction) as:

$$\Gamma^{\eta} = \left\{ \hat{y} \in Y \mid \sum_{i=1}^{k} I(\hat{y} \in \Gamma_i^{\varepsilon}) \geq \left\lceil \frac{k}{2} \right\rceil \right\} ,$$

where I is the indicator function; that is, Γ^{η} contains those labels which are contained in a majority of prediction sets.

We shall now establish the validity of such prediction set, under the assumption that the CPs are independent.

Theorem 2 (Ensemble of independent CPs is valid) *Consider the task of classifying a test object x with true label y. Let $C_1^{\varepsilon}, \dots, C_k^{\varepsilon}$ be k CPs, for some significance value ε. We assume the CPs are independent; i.e., considering their prediction sets Γ_i^{ε}, the events $\{y \notin \Gamma_i^{\varepsilon}\}$, $i = 1, \dots, k$, are independent. Then a majority vote ensemble produces an η-valid prediction set Γ^{η},*

$$Pr(y \notin \Gamma^{\eta}) = \eta$$

with

$$\eta = \sum_{i=0}^{\lfloor k/2 \rfloor} \binom{k}{i} \varepsilon^{k-i} (1-\varepsilon)^i .$$

Proof We derive the value of η analytically. The prediction sets $\Gamma_1^{\varepsilon}, \dots, \Gamma_k^{\varepsilon}$ define a binary vector $g := (g_1, \dots, g_k)$, where g_i is 0 if $y \notin \Gamma_i^{\varepsilon}$, 1 otherwise. The independence assumption on prediction sets means that g_i are drawn independently.

We call R the random variable counting the number of 1'es in g. Each g_i is a Bernoulli trial, with $Pr\,(g_i = 1) = 1 - \varepsilon$. Then R has a Binomial distribution, $R \sim B(k, 1-\varepsilon)$.

For y to be included in Γ^{η} we need at least $\lceil k/2 \rceil$ successes (i.e., $g_i = 1$), and the probability that y is *not* included is given by the CDF of $B(k, 1-\varepsilon)$:

$$Pr\left(y \notin \Gamma^{\eta}\right) = Pr\left(R \leq \left\lfloor \frac{k}{2} \right\rfloor\right) = \sum_{i=0}^{\lfloor k/2 \rfloor} \binom{k}{i} \varepsilon^{k-i} (1-\varepsilon)^i .$$

We set $\eta := Pr\,(y \notin \Gamma^{\eta})$. □

We will now determine how to compute the significance level ε that is required by each CP to obtain an η-valid ensemble. The following result shows that ε is obtained as a root of a k-th degree polynomial; one can efficiently solve such polynomial via numerical methods (e.g., using Newton's method, or by computing the eigenvalues of the companion matrix of the polynomial). In "Appendix", we provide a reference implementation.

This method for computing ε from η provides validity for independent CPs; we will refer to it as the *I-method*. In Sect. 4, we will develop the *C-method*, which guarantees conservative validity even for ensembles of correlated CPs.

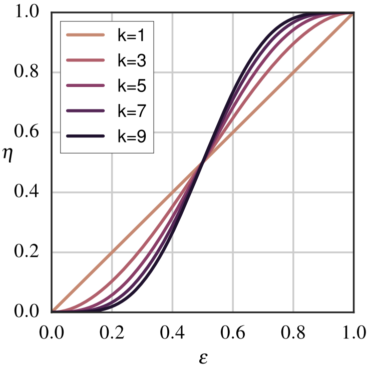

Fig. 1 Dependence of η on ε in the case of independent CPs

Proposition 1 [Determine ε for a desired η (*I-method*)] *Consider an η-valid majority vote ensemble of k independent CPs, $C_1^\varepsilon, \dots, C_k^\varepsilon$. The significance value ε required by each member of the ensemble to achieve η-validity is determined as the root $\varepsilon \in [0, 1]$ of the following polynomial:*

$$p(\varepsilon) = \sum_{j=\lceil k/2 \rceil}^{k} \alpha_j \varepsilon^j - \eta\,,$$

whose coefficients have the form:

$$\alpha_j := \sum_{i=k-j}^{\lfloor k/2 \rfloor} \binom{k}{i} \binom{i}{k-j} (-1)^{j-k+i} \quad j = \lceil k/2 \rceil, \dots, k\,.$$

Proof We rewrite $Pr\,(y \notin \Gamma^\eta)$ as a polynomial.

$$\begin{aligned} Pr\left(y \notin \Gamma^\eta\right) &= \sum_{i=0}^{\lfloor k/2 \rfloor} \binom{k}{i} \varepsilon^{k-i} (1-\varepsilon)^i \\ &= \sum_{i=0}^{\lfloor k/2 \rfloor} \sum_{j=0}^{i} \binom{k}{i} \binom{i}{j} (-1)^j \varepsilon^{k-i+j} \\ &= \sum_{i=0}^{\lfloor k/2 \rfloor} \sum_{j=k-i}^{k} \binom{k}{i} \binom{i}{j-k+i} (-1)^{j-k+i} \varepsilon^j \\ &= \sum_{j=k-\lfloor k/2 \rfloor}^{k} \sum_{i=k-j}^{\lfloor k/2 \rfloor} \binom{k}{i} \binom{i}{j-k+i} (-1)^{j-k+i} \varepsilon^j \\ &= \sum_{j=\lceil k/2 \rceil}^{k} \varepsilon^j \sum_{i=k-j}^{\lfloor k/2 \rfloor} \binom{k}{i} \binom{i}{k-j} (-1)^{j-k+i}\,. \end{aligned}$$

In the second step we expanded $(1-\varepsilon)^i$, in the third step we substituted j with $k-i+j$, and in the fourth step we used the fact that:

$$\sum_{i=0}^{a}\sum_{j=b-i}^{b}\beta_{ij}=\sum_{j=b-a}^{b}\sum_{i=b-j}^{a}\beta_{ij}\,;$$

finally, we used the equivalence $\binom{i}{i-j}=\binom{i}{j}$.

Given the coefficients, ε needs to satisfy the constraints: (i) be a solution to $p(\varepsilon) := Pr\,(y \notin \Gamma^{\eta}) - \eta = 0$, and ii) $\varepsilon \in [0, 1]$. □

We do not give formal proof of the fact that $p(\varepsilon)$ has a unique real root in the interval [0, 1]; it will suffice to observe that, for finite k and $\varepsilon \in [0, 1]$, $Pr\,(y \notin \Gamma^{\varepsilon})$ is strictly increasing (Fig. 1), and that $Pr\left(y \notin \Gamma^{0}\right) = 0$ and $Pr\left(y \notin \Gamma^{1}\right) = 1$, and thus, for a constant $\eta \in [0, 1]$, the polynomial $p(\varepsilon) = Pr\,(y \notin \Gamma^{\varepsilon}) - \eta$ has exactly one real root in [0, 1].

4 Beyond independence

So far, we assumed independence of the ensemble's CPs. We now discard this assumption, and reason about the validity of an ensemble of correlated CPs.

4.1 Measure of correlation

Correlation of classifiers depends on: (i) their structure, (ii) their training algorithm and hyper-parameters, and (iii) the data on which they are trained. Ultimately, such correlation reflects on the correlation of their predictions. One measure of dependence between two classifiers is the Q statistic (Yule 1900):

$$Q=\frac{N^{11}N^{00}-N^{01}N^{10}}{N^{11}N^{00}+N^{01}N^{10}}\,,$$

which can be computed on their predictions on a test set, where N^{11} is the count of objects for which both gave a correct prediction, N^{00} counts the objects which both misclassified, and N^{01} and N^{10} are respectively the number of objects for which one was correct, and the other one was not, and vice versa. Q takes values in $[-1, 1]$, where $Q = 0$ indicates independence. The correlation of an ensemble, Q_{av}, is the average Q statistic among its pairs of classifiers.

There is no optimal measure of correlation between classifiers. For this reason, in experiments, we will use the Q statistic merely as an indication to help interpreting the results. We compute Q on the output of two CPs as follows: we define an error as $I(y \notin \Gamma^{\varepsilon})$, and count the errors of each CP on a test set as required. To make Q_{av} independent of the significance level, we compute it for many values of $\varepsilon \in [0, 1]$, and average them. Future work may explore correlation measures accounting for the size of CPs' prediction sets.

4.2 Error bounds on correlated ensemble

In a thorough study on correlation and majority vote ensembles, Kuncheva et al. (2003) showed via synthetic examples that the accuracy resulting when combining k classifiers, each guaranteeing on a probability of error ε, is not related to their pairwise correlation in any trivial sense. Crucially, they showed that independence of classifiers in an ensemble is not necessarily a desideratum; in fact, a negative correlation is generally to be preferred, but also a positive correlation can be beneficial.

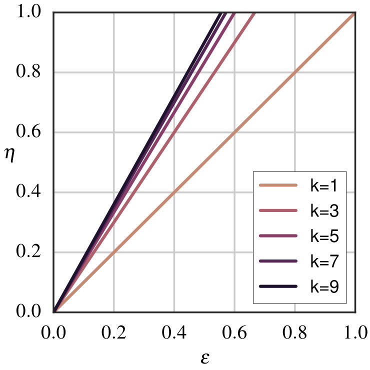

Fig. 2 Dependence of η on ε for the *C-method*

Kuncheva et al. (2003) also derived upper and lower error bounds for an ensemble of correlated classifiers, in the following sense. They defined the most favorable distribution on the classifiers' outputs ("pattern of success"), as a tendency of either exactly $\lfloor k/2 \rfloor + 1$ of them to be correct, or all of them being incorrect. They also defined a counter part of this distribution, the "pattern of failure", where classifiers can either be all correct, or exactly $\lfloor k/2 \rfloor$ correct and $\lfloor k/2 \rfloor + 1$ incorrect. Intuitively, an ensemble in the pattern of success makes the best use of each classifier's individual accuracy; conversely, under the pattern of failure, it wastes most correct votes. Assuming these are best and worst-case scenarios, they determined upper and lower bounds on the ensemble's error.

Because we are interested in proving validity for a CP ensemble, we will focus our attention on the upper bound of error, which is as follows.

Theorem 3 [η upper bound (Kuncheva et al. 2003)] *In the worst-case scenario ("pattern of failure"), the error η of an ensemble of odd k possibly correlated classifiers, each one guaranteeing on error probability ε, is:*

$$\eta = 1 - \max\left\{\frac{(1-\varepsilon)k - \lfloor k/2 \rfloor}{\lfloor k/2 \rfloor + 1}, 0\right\}$$

From this result, we now calculate the value ε required to guarantee η conservative validity for an ensemble of correlated CPs. We call *C-method* this way of obtaining ε from η.

Proposition 2 [Determine ε for a desired η (*C-method*)] *Let $C_1^\varepsilon, \ldots, C_k^\varepsilon$ be an ensemble of odd k CPs, where, for a desired significance level η, ε is determined as:*

$$\varepsilon = \eta \frac{\lceil k/2 \rceil}{k}\,. \tag{1}$$

Then the ensemble guarantees η conservative validity on its prediction set Γ^η for a new object with label y, i.e.:

$$Pr(y \notin \Gamma^\eta) \leq \eta\,.$$

This expression follows from Theorem 3 by performing simple calculations and noticing that $\lceil k/2 \rceil - \lfloor k/2 \rfloor = 1$ for odd k.

Figure 2 shows the relation between η and ε. We observe that the ε needed to obtain a certain η under this expression is much smaller than the one that was required in the independent case (Fig. 1). We remark, however, that ε as computed here takes into account the worst-case scenario, and that in practice the empirical error tends to be largely inferior to η (Sect. 5.3).

5 Empirical evaluation

We summarize our experiments as follows. Section 5.2 evaluates the validity of the *I-method*, for independent (Sect. 5.2) and correlated (Sect. 5.2) CPs; results confirm validity is guaranteed for ensembles with little or no correlation. Section 5.3 asserts the conservative validity of the *C-method* for correlated CPs. Section 5.4 evaluates efficiency improvements introduced by a CP ensemble. Finally, Sect. 5.5 compares the validity of the *I-method* with p-values combining methods.

5.1 Methodology

Data

We consider the following publicly available datasets.

`digits` (Pedregosa et al. 2011; Dheeru and Karra Taniskidou 2017)

Labels 10
Features 8 × 8 BW values
Examples 1797 (455 used for hyper-parameters selection)
Task Classify low resolution black and white images of digits.

`cifar-100-coarse` (Krizhevsky and Hinton 2009)

Labels 20
Features 32 × 32 RGB values
Examples 60K (10K used for hyper-parameters selection)
Task Classify 32 × 32 color images. We use the "coarse" version of the data, where the original 100 classes are grouped into 20 super-classes.

On the `digits` dataset, we verify the validity of *I-method* and *C-method*, and compare them with those suggested by Toccaceli and Gammerman (2017); we evaluate their efficiency on both `digits` and `cifar-100-coarse`.

Nonconformity measures

A CP is defined for a nonconformity measure. We will use the following nonconformity measures: k-NN, SVM (RBF kernel), decision trees, and random forest, each from the `scikit-learn`[3] implementation; details are in "Appendix".

For each experiment, we first perform a randomized grid search on a subset of data to choose good hyper-parameters for each nonconformity measure. Specifically, for k-NN we select $k \in \{1, 51, \ldots, 501\}$; for SVM, we select $\gamma \in \{10^{-9}, 10^{-8}, \ldots, 10^{3}\}$ and $C \in \{10^{-2}, 10^{-1}, \ldots, 10^{10}\}$; for both decision trees and random forest we select the minimum number of examples to perform a split from $\{10, 20, \ldots, 100\}$; for random forest we also select the number of estimators from the set $\{10, \ldots, 100\}$. The remaining hyper-parameters for each method are left to `scikit-learn`'s default ones.

[3] http://scikit-learn.org (Pedregosa et al. 2011).

Procedure

We compute the predictions of a CP ensemble in an on-line setting, by considering an increasing number of examples $n = 5, 6, \ldots$, up to the size of the dataset. For each n, and for a desired significance level η, we construct an ensemble of k CPs, $C_1^{\varepsilon}, \ldots, C_k^{\varepsilon}$, where ε is chosen according to the *I-method* or the *C-method*; we train each C_i^{ε} on the previous $1, 2, \ldots, n-1$ examples, and use it to make a prediction Γ_i^{ε} for the n-th example; finally, we construct the ensemble's prediction set Γ^{η} for this example by taking a majority vote for each label across the CPs' prediction sets Γ_i^{ε}, as shown in Sect. 3.

5.2 Validity of the *I-method*

We evaluate validity of the *I-method* by counting the errors committed and then plotting them against the expected error for a given significance level η.

Independent CPs

We first consider the case of independent or low correlated CPs, where the *I-method* for determining ε for a desired η guarantees exact validity.

In Sect. 4, we indicated the possible reasons of correlation between classifiers. An elegant way for obtaining k classifiers with low inter-correlation is to train each on a separate subset of features. For example, to achieve this on the `digits` dataset, we divide its 64 features into $k = 5$ sets of approximately 13 features each, and then we train each CP on one of them. A similar technique was used by Kuncheva and Whitaker (2003). To further reduce the correlation, we select a different nonconformity measure for each CP in the ensemble.

We refer to an ensemble with the nonconformity measures of its CPs. We evaluate the following ensembles, two of which with $k = 3$ components, one with $k = 5$ components. Each CP is trained on a separate subset of features.

Ensemble	Nonconformity measures	Q_{av}
`DT+LR+SVM`	Decision tree, logistic regression, SVM	– 0.11
`k-NN+LR+SVM`	k-NN, logistic regression, SVM	0.06
`DT+LR+SVM+k-NN+RF`	Decision tree, logistic regression, SVM, k-NN, random forest	0.13

In this table, Q_{av} is the Q statistic correlation measure, averaged among an ensemble's CPs and for many values of ε as shown in Sect. 4.1; Q_{av} takes values between -1 and $+1$, where 0 indicates independence. We remark that an ensemble with negative correlation can reduce the expected error η, while correlation close to 0 will guarantee an error close to η.

All the ensembles have a low correlation; specifically, `DT+LR+SVM` has negative dependence, while the others present a slightly higher Q_{av}. Notably, the ensemble with the highest correlation is also the largest; this may suggest that, as the size of an ensemble increases, its classifiers are more likely correlated.

We compute the cumulative empirical error of each ensemble on the `digits` dataset in an on-line setting, for many values of η; here, the significance level ε of the individual CPs is determined using the *I-method*. Figure 3 indicates that the cumulative error of all

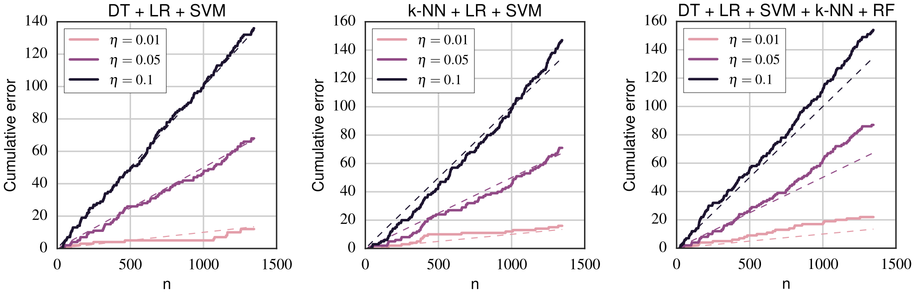

Fig. 3 Empirical validity of the *I-method* for CP ensembles under low correlation conditions, on the `digits` dataset

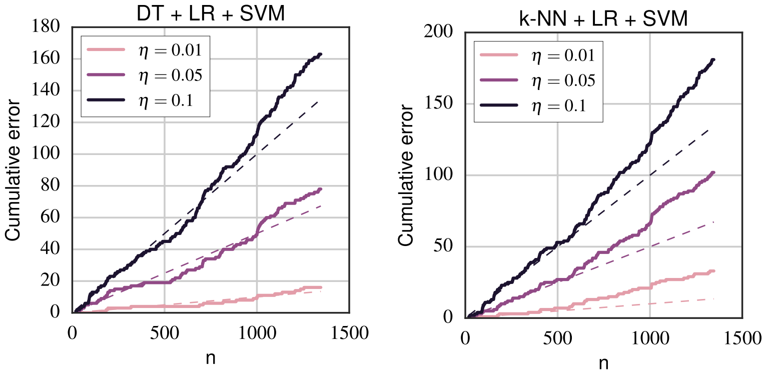

Fig. 4 Cumulative error of two ensembles on the `digits` dataset. Validity of *I-method* is significantly violated when the ensemble's CPs are strongly correlated (right)

ensembles is close to ηn, which means validity is achieved. We notice, however, that the error of `DT+LR+SVM+k-NN+RF` is often higher than η, probably due to its larger correlation.

Correlated CPs

We now evaluate the validity of the *I-method* for CP ensembles with higher correlation. To this end, we train each CP on the full set of features. We remark that, in this case, the *I-method* is not expected to guarantee validity.

For this experiment, we consider two ensembles: `DT+LR+SVM`, `k-NN+LR+SVM`. Under these conditions, the correlation Q_{av} of `DT+LR+SVM` is 0.12, and 0.25 for `k-NN+LR+SVM`; while the former maintained a correlation close to 0 with respect to our previous experiment, the correlation of the latter increased notably.

Figure 4 reports the error of the ensembles on the `digits` dataset. Here, validity is obtained for the first ensemble, while it is violated by the second one; this is not surprising, given what we observed in terms of their correlation.

5.3 Validity of the *C-method*

We verify the conservative validity (i.e., error is always smaller or equal to η) of the *C-method* for an ensemble of correlated CPs.

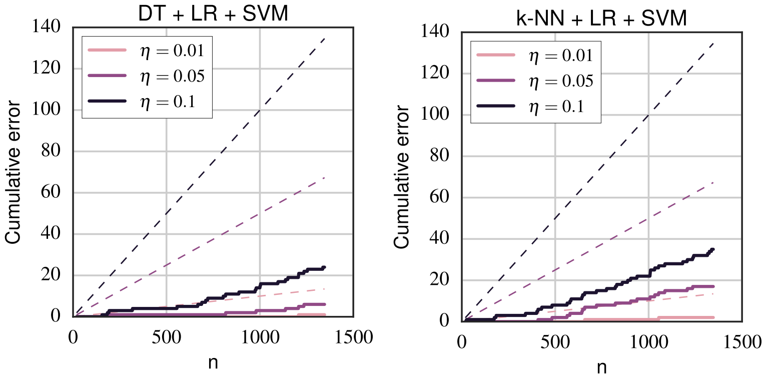

Fig. 5 Empirical conservative validity of the *C-method* for ensembles of correlated CPs, on the `digits` dataset

We consider the two ensembles used in the previous experiment, but this time we compute ε from η using the *C-method*. Results in Fig. 5 indicate that conservative validity is satisfied. We also notice that the empirical error is much smaller than the chosen significance level η; indeed, η is now an upper bound, which will only be reached in practice under unlucky (and unlikely) circumstances (Sect. 4.2).

5.4 Efficiency

We measure an ensemble' efficiency as the average size of its prediction sets (Sect. 2). We evaluate on the `digits` dataset all the ensembles considered so far. We also assess the following ensemble on the `cifar-100-coarse` dataset: `k-NN+RF+DT`, composed of k-NN, random forest, and decision tree.[4] Ensembles' correlation Q_{av} is: 0.12 (`DT+LR+SVM`), 0.25 (`k-NN+LR+SVM`), 0.49 (`k-NN+RF+DT`).

Results are shown in Table 1. We observe that the *I-method* tends to produce more efficient predictions than its member CPs; in particular, the `k-NN+LR+SVM` outperforms each of its CPs. The price one has to pay, in this case, is that validity is slightly violated because of the correlation.

The *C-method* produces wider predictions than the *I-method*, and the efficiency of its ensembles only outperforms the worst efficient of their CPs. However, this method achieves an empirical error that is much lower than the significance level η.

5.5 Validity comparison with Fisher's and Stouffer's *p*-values combining

Balasubramanian et al. (2015) and Toccaceli and Gammerman (2017) independently proposed ensembles by combining CPs' p-values with Fisher's and Stouffer's methods, both of which are valid for independent ensembles.

We compare experimentally the deviation from validity of these methods with the *I-method*, for ensembles of correlated CPs; under these circumstances, all these methods' validity should be violated. We consider ensembles trained as in Sect. 5.2, on the `digits` dataset, for a significance level $\eta = 0.1$.

[4] Because of the computational cost of CP, on the `cifar-100-coarse` dataset we used Inductive CP, an approximation of CP that was shown to guarantee validity in practice.

Table 1 Efficiency of individual CPs and respective ensembles using *I-method* and *C-method*

`digits` dataset

					DT+LR+SVM		k-NN+LR+SVM	
η	DT	k-NN	LR	SVM	*I-method*	*C-method*	*I-method*	*C-method*
0.01	7.97	1.95	2.73	9.88	4.80 (0.01)	8.51 (0.00)	1.75 (0.02)	3.61 (0.00)
0.05	5.42	1.42	1.66	9.46	3.20 (0.06)	6.41 (0.00)	1.28 (0.08)	2.10 (0.01)
0.1	3.75	1.22	1.25	8.65	2.63 (0.12)	4.35 (0.02)	1.10 (0.14)	1.68 (0.03)

`cifar-100-coarse` dataset

				k-NN+RF+DT	
η	k-NN	RF	DT	*I-method*	*C-method*
0.01	19.17	18.59	19.73	17.66 (0.03)	19.63 (0.00)
0.05	17.40	16.29	18.66	14.99 (0.09)	18.57 (0.01)
0.1	15.46	14.26	17.30	12.92 (0.14)	17.37 (0.03)

Lower values mean less uncertain predictions. Empirical error of the ensembles is in parentheses; the error of individual CPs is η

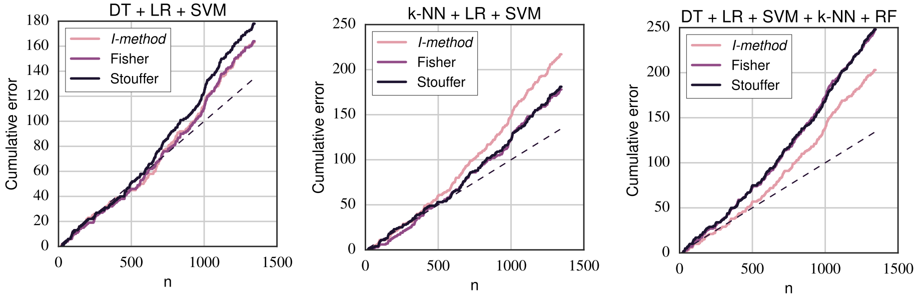

Fig. 6 Comparison of the *I-method* with Fisher's and Stouffer's p-values combining methods on the `digits` dataset for ensembles of correlated CPs

Figure 6 indicates no clear winner: each method is closer to validity in some experiments, but it fails in others. Similarly, the average prediction set size of all methods is close to 1 for the last two ensembles; however, for `DT+LR+SVM`, the *I-method* is less efficient (average $|\Gamma^\eta| = 2.63$) than Fisher's and Stouffer's methods (attaining respectively 1.34 and 1.69).

6 Conclusions and future work

There exist two strategies for forming CP ensembles: (i) combining p-values and thresholding them for a desired significance level η (Toccaceli and Gammerman 2017; Linusson et al. 2017), or (ii) taking a vote (e.g., majority) on ε-valid prediction sets, where ε depends on a desired significance level η. This paper considered the latter, and derived exact validity guarantees for majority vote ensembles of independent CPs (*I-method*), and conservative validity for ensembles of correlated CPs (*C-method*). We previously used the technique of adapting ε to combine prediction sets and obtain η-validity in the context of Hidden Markov Models (Cherubin and Nouretdinov 2016).

The greatest theoretical advantage of majority vote ensembles is that they ease proving guarantees for correlated CPs (e.g., see *C-method*), and, in general, they inherit properties from the vast literature on majority vote classification. Proving similar results for p-values combining methods is a much harder task; e.g., Linusson et al. (2017) could not prove validity for arbitrary nonconformity measures, and the validity guarantees achieved in the work by Balasubramanian et al. (2015), Toccaceli and Gammerman (2017) only held for independent CPs.

Conveniently, p-values combining methods output a p-value as intermediate value, which is more informative than the output of a majority vote ensemble (i.e., a prediction set). We remark, however, that if an application needs predictions associated with a confidence measure, one can use a majority vote ensemble for prediction, and combine the p-values using p-values combining methods, with minor impact on computational costs.

In computational terms, the time complexity of combining the predictions of an ensemble of k CPs for N test objects is $O(kLN)$, where $L = |Y|$, for both majority vote and p-values combining methods; this is negligible with respect to the complexity of generating the CP predictions. Majority vote, however, has simpler operations to execute for each step, so it will generally be faster than combining p-values using Fisher's or Stouffer's methods. On the other hand, if one wanted to generate prediction sets for many significance levels η from a majority vote ensemble, they would have to re-combine predictions; this is not needed for p-values combining methods.

We showed that the *C-method*, whilst guaranteeing conservative validity, tends to attain a much smaller error than η; on the other hand, the *I-method* achieved better efficiency, with minor effects on the validity. Future work may compromise between the *I-method* and the *C-method* to balance their advantages (e.g., by averaging ε obtained by the two methods); while this approach would not generally give formal guarantees, it may be able to reach the desired balance in practice.

Kuncheva et al. (2003) elaborated on how η and ε depend on the correlation Q in the best and worst-case scenarios. However, they also showed that, in general, there is no precise relationship between an ensemble's probability of error η and its correlation. An interesting line of research is to investigate further the dependency between these variables; in particular, correlation measures for CPs could exploit further information from their prediction sets (e.g., average size), which may give heuristics for adapting ε to Q.

Acknowledgements I am grateful to Alexander Gammerman, Henrik Linusson, Paolo Toccaceli, Fernando Virdia, and Vladimir Vovk for useful discussion and feedback. I thank the anonymous reviewers for their very careful reviews and insightful suggestions. The author was supported by the EPSRC and the UK government as part of the Centre for Doctoral Training in Cyber Security at Royal Holloway, University of London (EP/K035584/1).

Appendix A: Conformal prediction

A CP $C^{A,\varepsilon}$ accepts training examples $(x_i, y_i)_{i=1}^n$ and a new object x, and predicts a set of candidate labels $\Gamma^\varepsilon \subseteq Y$ for x. Let $z_i = (x_i, y_i), i = 1, \ldots, n$. A CP with nonconformity measure $A : (X, Y)^* \times (X, Y) \mapsto \mathbb{R}_{\geq 0}$ works as follows.

```
Function $C^{A,\varepsilon}(x, (z_1, \ldots, z_n))$:
    Initialize $\Gamma^\varepsilon$ to the empty set
    for $\hat{y} \in Y$ do
        Set temporarily $z_{n+1} = (x, \hat{y})$
        for $i = 1, \ldots, n+1$ do
            $\alpha_i \leftarrow A(z_i, (z_1, \ldots, z_{n+1}) \setminus z_i)$
        end
        $\tau \xleftarrow{\$} Uni(0, 1)$
        $p_{\hat{y}} \leftarrow \frac{\#\{i|\alpha_i > \alpha_{n+1}\} + \tau\#\{i|\alpha_i = \alpha_{n+1}\}}{n+1}$
        if $p_{\hat{y}} > \varepsilon$ then
            Add $\hat{y}$ to $\Gamma^\varepsilon$
        end
    end
return $\Gamma^\varepsilon$
```

We used nonconformity measures based on the following algorithms: k-NN, SVM, decision trees, and random forest. We constructed nonconformity measures from them using the *margin error* as follows (Johansson et al. 2013; Linusson 2015). Let $f : X \mapsto [0, 1]^L$, with $L = |Y|$, be a classifier trained on examples $(x_i, y_i)_{i=1}^n$ by using one of the learning algorithms above; the classifier outputs, for a test object x, a confidence vector $c = f(x)$, where $c_{\hat{y}} \in [0, 1]$ is the confidence value for label $\hat{y}$. Then, for an example (x, y), we compute a nonconformity score as follows: $1/2 - (c_y - \max_{y_i \neq y} c_{y_i})/2$.

Appendix B: Code of the *I-method*

Unoptimized `Python 2.7` reference implementation of the *I-method*.

```
import numpy as np
from math import floor, ceil, factorial

def imethod(k, eta):
    k2c = int(ceil(k/2.0))
    k2f = int(floor(k/2.0))
    # n choose k.
    choose = lambda n, k: factorial(n) / (factorial(k)*factorial(n-k))
    # Determine the coefficients of the polynomial.
    coefs = []
    for j in range(k2c, k+1):
        c = 0
        for i in range(k-j, k2f+1):
            c += choose(k, i) * choose(i, k-j) * (-1)**(j-k+i)
        coefs.append(c)
    # Set remaining coefficients to 0, and the first to -eta.
```

```
coefs = [-eta] + [0]*(k - len(coefs)) + coefs
# Find root in the interval [0,1].
for root in np.roots(coefs[::-1]):
    if root.imag != 0:
        continue
    if 0 <= root <= 1:
        return root.real
raise Exception("No root was found")
```

References

Balasubramanian, V. N., Chakraborty, S., & Panchanathan, S. (2015). Conformal predictions for information fusion. *Annals of Mathematics and Artificial Intelligence*, *74*(1–2), 45–65.

Cherubin. G., & Nouretdinov, I. (2016). Hidden markov models with confidence. In *Symposium on conformal and probabilistic prediction with applications*, (pp. 128–144), Springer.

Dheeru, D., & Karra Taniskidou, E. (2017). *UCI machine learning repository*. http://archive.ics.uci.edu/ml.

Johansson, U., Bostrom, H., & Lofstrom, T. (2013). Conformal prediction using decision trees. In *IEEE 13th international conference on data mining (ICDM), 2013 , IEEE*, (pp. 330–339).

Krizhevsky, A., & Hinton, G. (2009). *Learning multiple layers of features from tiny images*. Technical report, University of Toronto.

Kuncheva, L. I., & Whitaker, C. J. (2003). Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, *51*(2), 181–207.

Kuncheva, L. I., Whitaker, C. J., Shipp, C. A., & Duin, R. P. (2003). Limits on the majority vote accuracy in classifier fusion. *Pattern Analysis & Applications*, *6*(1), 22–31.

Linusson, H. (2015). *Nonconformist, python implementation of the conformal prediction framework*. https://github.com/donlnz/nonconformist.

Linusson, H., Norinder, U., Boström, H., Johansson, U., & Löfström, T. (2017). On the calibration of aggregated conformal predictors. In *Proceedings of machine learning research*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Toccaceli, P., & Gammerman, A. (2017). Combination of conformal predictors for classification. In *Proceedings of machine learning research*.

Vovk, V., Gammerman, A., & Shafer, G. (2005). *Algorithmic learning in a random world*. New York: Springer.

Vovk, V., Fedorova, V., Nouretdinov, I., & Gammerman, A. (2016). Criteria of efficiency for conformal prediction. In *Symposium on conformal and probabilistic prediction with applications*, (pp. 23–39), Springer.

Yule, G. (1900). On the association of attributes in statistics. *Philosophical Transactions of the Royal Society*, *194*, 257–319.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.